ACCELERATED FAILURE-TIME MODEL WITH WEIGHTED LEAST-SQUARES ESTIMATION: APPLICATION ON CARDIOVASCULAR DISEASE

Enoch Yabkwa Yanshak[1], Yakubu Aliyu[1], Sani Ibrahim Doguwa[1], Kolawole Daramola[1], Emmanuel Lekwot[1]

[1]*Department of Statistics, Faculty of Physical Sciences, Ahmadu Bello University, Zaria, Nigeria*

ABSTRACT

In the study we present the comparison among standard accelerated failure time models and weighted least square estimation. Data were collected from Barau Dikko Teaching hospital Kaduna-Nigeria. The standard accelerated failure-time model is a more powerful and interpretable model than the Cox proportional hazards model, provided that model imposed distribution and homoscedasticity assumptions satisfied. However, most of the real data are heteroscedastic which violates the fundamental assumption and consequently, the statistical inference could be erroneous in accelerated failure-time modelling. The weighted least-squares estimation for the accelerated failure-time model is an efficient semi-parametric approach for time-to-event data without the homoscedasticity assumption, which is developed recently and not often utilized for real data analysis. Thus, this study was conducted to ascertain the better performance of the weighted least-squares estimation method over the standard methods. We analysed a REAL dataset of cardiovascular diseases patients we recently collected. We compared the results from standard methods of estimation for the accelerated failure-time model with the results revealed from the weighted least-squares estimation. We found that the data are heteroscedastic and indicated that the weighted least-square method should be used to analyse this data. The weighted least-squares estimation revealed more accurate, and efficient estimates of covariates effect since its confidence intervals were shorter and it identified more significant covariates with smaller AIC. Accordingly, the survival of cardiovascular patients was found to be significantly linked with age, over weight, body mass index, diabetes mellitus, irregular pulse rate, alcoholic usage, smoked serum-creatinine, ejection-fraction, anamia and seum-sodium. The weighted least-squares estimation performed the best in providing more significant effects and precise estimates than the standard accelerated failure-time methods of estimation if data are heteroscedastic.

Keywords: Accelerated Failure Time Models, Weighted Least Square Estimation, Survival function, Heteroscedastic, Hazard –Function, Cox Proportional Model.

1. INTRODUCTION

Survival analysis plays a crucial role in comprehending the timing of events, such as death or disease recurrence, in various medical domains. In the context of cardiovascular patients, survival analysis offers valuable insights into the factors influencing the time to survival and aids in identifying significant risk factors related to cardiovascular events (Spoto, *et al.* 2021). There are numerous survival models available for analysing such data, each with its own strengths and underlying assumptions.

The Accelerated Failure Time (AFT) Model establishes a linear association between the logarithm of survival time and the covariates, assuming that variables' effects on survival time are multiplicative

(Swindle & Mukhopadhyay, 2011). The AFT model comes with strong assumptions about the distribution of event time, it is relatively easier to implement and interpret. It performs well when the underlying distribution assumptions are met, and it does not require assessing the assumption of proportional hazards, beforehand, unlike the Proportional Hazards (PH) models, in AFT models, survival data are not assumed to have constant hazards, making them an alternative parametric approach to commonly use in place of proportional hazard models. They are widely employed in modern medical statistics, actuarial works Richard, (212), and various applications, including retention analysis.

1.1 Problem Statement

Cardiovascular diseases remain a significant global health challenge, necessitating accurate prediction of survival outcomes and the identification of influential risk factors to improve treatment and patient care. Parametric survival models are widely used in medical research for time-to-event data analysis. In Nigeria, researchers have applied the Accelerated Failure Time (AFT) model to study survival times in diseases such as liver cirrhosis, breast cancer, lung cancer, kidney transplant, and obstetric fistula. However, its application to cardiovascular disease data remains unexplored. Victor et al (2017) identified critical risk factors for cardiovascular diseases, Including age, gender, smoking status, diabetes mellitus, comorbidities, and treatment modalities. Similarly, Barnwal and Hocking, (2022) conducted a comparative study on survival analysis for obstetric fistula patients in Nigeria. Despite these Contributions, there is a need to determine the most suitable survival model for examining factors associated with the Survival time of cardiovascular patients. Furthermore, the use of weighted least squares estimation for cardiovascular data, Particularly when the assumption of constant variance is violated, has not been Explored. By comparing various survival Models and applying weighted least squares Estimation, this study aims to Comprehensively assess their performance in Predicting survival outcomes for Cardiovascular patients.

1.2 Objectives

The objectives of this study are to identify the best-fitting Accelerated Failure-Time (AFT) model for cardiovascular disease data, evaluate the validity of the constant variance assumption, and apply Weighted Least-Squares (WLS) estimation if this assumption is violated. Additionally, the study aims to assess key risk factors and explore factors influencing the survival time of cardiovascular patients, using data from Kaduna State, Nigeria.

2. LITERATURE REVIEW

Tanvir, *et al.* (2017) Analysed cardiac arrest patients in Pakistan aged 40+ with left ventricular systolic dysfunction in YHA classes III-IV. Focused on hospitalized patients During April-December 2015. Victor *et al.* (2017) Investigated cardiovascular disease (CVD) risk factors in a Nigerian population with impaired fasting glucose (IFG) and diabetes. Found a higher prevalence of co-morbidities in IFG patients compared to diabetic individuals, with dyslipidemia being the most common, utilized Optimal Discriminant and HO-CTA techniques. Belaynesh and Zeytu, (2021). Conducted a survival analysis of cardiac patients in Ethiopia using parametric, semi-parametric, and non-parametric models. The Weibull AFT model was most suitable, revealing male patients had 1.9 times higher mortality risk. Alvaro and René, (2021) Examined heart failure survival in Ecuador, identifying an overall 5-year survival rate of 46%. Age and heart failure etiology were significant prognostic factors.

In addition Abtsega, *et al.* (2019) analysed survival factors in Ethiopian hypertension patients using Weibull and other models. Significant factors included age, residence, family history, and treatment adherence. Edward, *et al.* (2015). Compared statistical methods for analysing cardiovascular risks, highlighting differences in hazard ratios when using models for recurrent and multiple event types. Rana and Shuaib, (2006) Compared hazard modeling distributions, finding Weibull, Erlang, and Gamma distributions effective for approximating lifetimes, while the lognormal distribution was less precise for extended times. Hui, *et al.* (2011) Evaluated prognostic factors for gastric cancer using Weibull and Cox models, finding the Weibull model more precise based on Akaike Information Criterion (AIC). Yu and Liu (2018) Proposed a homoscedasticity test for the AFT model, emphasizing its importance in maintaining accurate survival analysis conclusions. Li *et al.* (2020) Demonstrated Weibull regression modelling using R, highlighting techniques for evaluating model fit and presenting findings visually. Zhongheng, (2016) Found modified Weibull models superior for analysing coronary heart disease survival compared to traditional Weibull models. Yesuf, and Ding-Geng, (2021) Applied weighted least-squares estimation in AFT models for HIV-positive patient survival, finding it more effective than classical methods for heteroscedastic data. Piotr, and Aaron, (2022) Introduced scalable techniques for high-dimensional AFT models, showing improved performance and efficiency in penalized rank-based estimation.

3. METHODOLOGY

The current investigation relied on secondary data sourced from the records and information sheets of cardiovascular patients treated at Barau Dikko Teaching Hospital in Kaduna state. The Focus was on patients who underwent cardiac procedures either before or after operations and were being monitored at the hospital from January 2014 to December 2022. Various factors potentially linked to the mortality of cardiovascular patients were scrutinized, with the average time to survival after admission serving as the primary variable of interest. This time span represents the period from admission for treatment to either the date of death or censoring. Censoring applies to patients who remain alive throughout the study or are lost to follow-up before experiencing the event of concern (death).

Drawing from extensive literature reviews, insights from seasoned researchers, and the researcher's own expertise, several explanatory variables were taken into account. These variables include age, gender, hypertension, body mass index, smoking habits, alcohol consumption, diabetes, ejection fraction, serum creatinine levels, anemia, pulse rate, geographical region, creatinine phosphokinase levels, and patient status. The data collected underwent analysis using the Python programming language.

3.1 The Accelerated Failure Time Models

The accelerated failure time models sometimes referred to as the accelerated life models create a linear relationship between the logarithm of the failure time and independent variables (Fox, 2020). For doing regression analysis on censored failure time data, this model provides an appealing substitute to the popular Cox (1972) proportional hazards model due to its simple physical explanation. In the accelerated failure time model, as opposed to the PH model, we examine the direct impact of explanatory factors on the survival time.

The hazard function of accelerated failure time model is expressed as:

$$h(t/x) = h_0[t\exp(-X^{'}\beta^{*})]\exp(-X^{'}\beta^{*}) \qquad (1)$$

The survival function is given as:

$$s(t/x) = \exp\{-H_0[t\exp(-X^{'}\beta^{*})]\} \qquad (2)$$

The probability density function is given a

$$f(t;x) = h_0[t\exp(-X^{'}\beta^{*})]\exp(-X^{'}\beta^{*})\exp\{-H_0[t\exp(-X^{'}\beta^{*})]\} \qquad (3)$$

The time scale in AFT regression models is determined by the impact of variables in a technique that if $\exp(X^{i}\beta^{*}) > 1$ the covariate vector has the effect of slowing the survival process, and if the $\exp(X^{'}\beta^{*}) < 1$. The effect is to accelerate it Barnwal *et al.* (2022), where $X^{'}$ is a covariate vector $h_o(t)$ is referred to baseline hazard. $\exp(X^{'}\beta)$ is multiplicative effect term.

3.2 The Exponential Accelerated Failure Time Regression Model

The exponential failure time model is a statistical model commonly used in reliability and survival analysis. It describes the time until an event of interest (such as failure or death) occurs for a particular unit or system. The key assumption of this model is that the hazard rate, which represents the instantaneous failure rate at any given time, is constant over time (Yu and Liu 2018; Austin, *et al.* 2021). The hazard function is given as:

$$h(t/x) = h_0[\exp(X^{'}\beta)] \qquad (4)$$

Survival function is expressed as:

$$S(t:x,\beta^{*}) = \exp[-\exp(y - X^{'}\beta^{*})] \qquad (5)$$

In terms of the extreme value distribution, which is provided by, the density function may be stated as:

$$f(t;x,\beta^{*}) = \exp[(y - X^{'}\beta^{*}) - \exp(y - X^{'}\beta^{*})] \qquad (6)$$

3.3 Weibull Accelerated Failure Time Regression Model

In this section, the idea of Yusuf and Ding-G-Geng, (2021) has been applied using Weibull Accelerated Failure Time Regression Model. Thus, Survival function of weibull accelerated failure time model may be represented as;

$$s(t;x,\beta^{*},p^{*}) = \exp[-\exp(\frac{y - X^{'}\beta^{*}}{p^{*}})] - \infty < y < \infty \qquad (7)$$

p^{*} is the scale parameter

the Weibull hazard function as an inverse in term of the AFT;

$$h(t:x,\beta^*,p^*)=(p^*)^{-}\exp(\frac{y-X'\beta^*}{p^*}) \quad -\infty<y<\infty \tag{8}$$

The AFT density functions of the Weibull regression model may be directly stated as:

$$f(t;x,\beta^*P^*)=\exp[-\exp(\frac{y-X'\beta^*}{p^*})](p^*)^{-1}\exp(\frac{y-X'\beta^*}{p^*})(p^*)^{-1}\exp[(\frac{y-X'\beta^*}{p^*})- \exp(\frac{y-X'\beta^*}{p^*})],-\infty<y<\infty \tag{9}$$

3.4 Lognormal Accelerated Failure Time Regression Model

To describe a monotonic risk function process, the lognormal distribution often used parametric function. Since the logarithm of lognormal distribution is utilized when assuming that AFT survival times follow a log-normal distribution, it is simple, which contributes to its wide usage (Javeria *et al.*, 2021; Piotr & Aaron 2022). The baseline survival function and hazard function are provided by:

$$S_0(t)=1-\phi\left(\frac{\log t-\mu}{\sigma}\right) \tag{10}$$

$$h_0(t)=\frac{\phi\left(\frac{\log t}{\sigma}\right)}{\left[1-\phi\left(\frac{\log t}{\sigma}\right)\right]\sigma t} \tag{11}$$

μ is the intercept, σ is the scale parameter and X_i is a random variable, $\phi(x)$ is the cumulative density function of the standard normal distribution. The ith individual's survival function is

$$S_i(t)=s_0\left(\frac{t}{\eta_i}\right)=1-\phi(\frac{\log t-\alpha^i x_i}{\sigma}) \tag{12}$$

when $\eta i = \exp(\alpha_1 x_1 + \alpha_2 x_2 + \cdots + \alpha_p x_p)$. Consequently, i^{th} individual's log survival time has been normal $(\mu+\alpha^I x_i,\sigma)$. The AFT property applies t

3.5 General View of Weighted Least Square Estimation

Here we applied the idea of (Yu and Liu 2018; Qingkai, 2023) when the assumption of standard method is violated. Thus, the general view of weighted least square estimation is given as:

$$MSE(b)=\frac{1}{n}\sum_{i=1}^{n}(Y_i-X_i\beta)^2 \tag{13}$$

where X_i is the i^{th} row of X. Weighted MSE is given as:

$$WMSE(b, w_1, ...w_n) = \frac{1}{n}\sum_{i=1}^{n} w_i (Y_i - X_i b)^2 \tag{14}$$

3.6 Heteroskedasticity

As also applied by (Yu and Liu, 2018); Yesuf, and Ding-Geng, 2021)

$$Y_i = \beta_0 + \beta_1 X_{1i} + ... + \beta_p X_{pi} + e_i$$

where $\mathrm{E}[e_i] = 0$ *and* $Var[e_i] = \sigma_i^2$.(As usual, we are treating the X_i's as fixed). This is called the *Heteroskedastic linear regression model*. For now assume we know $\sigma_1, ..., \sigma_p$.

The model is

$$Y = X\beta + e \tag{15}$$

3.7 Accelerated Failure Time Model With Weighted Least Square Estimation

The AFT approach is superior in many respects (Yu & Liu, 2018). Traditionally, the AFT model is inferred using the rank and least-squares approaches (Orbe & Núñez-Antón, 2006). By adopting a log-linear form with a certain error's survival time distribution component, the traditional AFT models as a whole are brought together.

Weighted least-squares analysis is one of the most useful ways for analyzing real data that violate the homoscedasticity assumption. According to Yu, and Liu (2013), the WLSE employs a weighted least-squares equation with synthetic observations that are weighted by the square root of their variances, with the variances being computed using the local polynomial regression. Thus, the weighted regression according to Yu, and Liu (2013) is as follows.

$$T_{inew}^* = \alpha_0 x_{i0new} + \beta_0^T X_{inew} + e_i, \text{ i=1,2,…,n} \tag{16}$$

where: $T_{inew}^* = T_i^* / \sigma NPE(\mu_i); \sigma NPE(\mu_i)$ s the non-parametric estimator (NE) via the local polynomial regression of Based on the weighted least-squares regression in equation (16) all ~ β₀, including α₀, are slope parameters.

3.8 Model Selection

The proportional hazard model is the model that is most frequently used to model survival data. The choice of the fitted model will be made using the Bayesian information criteria (BIC) (Aerts *et al.*, 2010)and the Akaike Information Criterion (AIC) by Akaikes (2023). The model with the lowest coefficients would be the one that best fits the data.

3.9 Akaike Information Criterion (AIC)

The model with the lowest AIC score appears to have the best fit when comparing models that are run with various parametric forms. While these tools are useful for evaluating how well the stated form fits the data, the most important consideration when selecting a parametric form is always the plausibility of the indicated underlying danger. Once it is determined that the selected parametric form successfully fits the data, comparisons between other models can be made using techniques similar to those previously discussed for semi-proportional hazard models, such as residual plots and goodness-of-fit tests.

The selected model is anticipated to reduce the difference between the model and the real world data. The model with the lowest AIC is regarded as the best fit when numerous competing models are given a dataset and ranked according to their individual AIC values. Hirotogu Akaike first proposed the AIC in 1973 as an expansion of the maximum likelihood concept. The formula is a mathematical representation of the AIC

$$AIC = -2ll + 2(p + k) \tag{17}$$

where p is the number of variables and ll is the log-likelihood (the probability of the data in a certain model), or number of parameters, and *k (constant factor), k =1* for exponential model and *k =2* for Weibull and Log-logistic model (Klein *et al.*, 2005). Smaller AIC indicates better model.

3.10 Bayesian Information Criteria (BIC)

The Bayesian Information criteria (BIC), commonly referred to as the Schwarz criteria, is another statistical metric used for comparing and evaluating time series models. It serves as a means of evaluating the quality of fit of candidate models and aims to identify the true model that best represents the data. The model with the minimum BIC value is considered the true model and is considered the best fit for the data. Gideon Schwarz, a statistician, developed this criterion, and it is closely related to AIC. The key difference between AIC and BIC becomes apparent when we add a number of K parameters to enhance the model's goodness of fit. In such cases, BIC penalizes more compared to AIC and exhibits consistency in model selection. As the sample size grows, BIC is guaranteed to select the true model among the candidate models considered, as long as the true model is part of those candidates. It also favours more parsimonious model than complex model, and it is use to access two competing model (i.e. when more than one model is true)

$$BIC = -2ll + 2k\log(n) \tag{18}$$

where ll is the model log-likelihood function, n is the number of observation and K is the number of parameter. It can also be written as

$$BIC = -2\ln\{p(x/\hat{\theta})\} + k\ln n \tag{19}$$

where x is the random variable, $\hat{\theta}$ is the maximum likelihood estimate, k is the number of parameter and n is the sample size.

3.11 Sampling Techniques and Sample Size Determination

In the present study, a representative sample of cardiac patients will be selected using a simple random sampling technique from a large pool of patients. A sample size of 300 patients was chosen for the research. The models that will be used in this research have been obtained.

4. RESULTS AND DISCUSSIONS

The response variable of interest for this study was survival time of cardiovascular patients until death in days and the independent variables considered are in two groups, the continuous group and the categorized group. The continuous variables are Age, serum_sodium, ejection_fraction, Creatinine_phosphokinase and serum_creatinine. The categorized variables are anemia, diabetes mellitus, alcoholic usage, smoking, high blood pressure, body mass index, sex, and region and pulse rate.

In this section, we delve into the analysis and discussion. The data was obtained from Barau Dikko Teaching Hospital Kaduna.

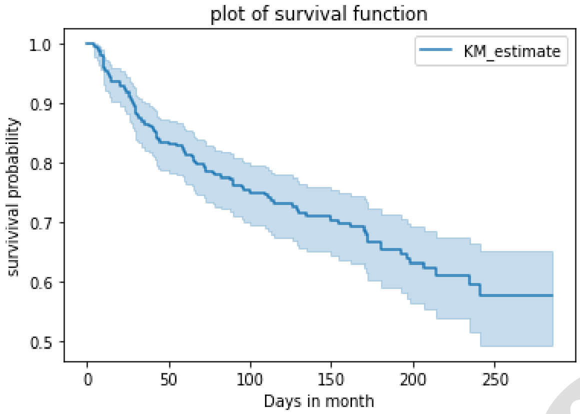

Figure 1: KMF Plot

Figure1 illustrates how the survival probabilities changes over the time horizon. As time passes, the Survival probabilities of cardiovascular reduce the response variable of interest for this study was survival time of cardiovascular patients until death in days and the independent variables considered are in two groups, the continuous group and the categorized group. The continuous variables are Age, serum_sodium, ejection_fraction, Creatinine_phosphokinase and serum_creatinine. The categorized variables are anemia, Diabetes mellitus, alcoholic usage, smoking, high blood pressure, body mass index, sex, region and pulse rate .

Table 1: Present Selection of best fitted model

Model Type	Model	Obs	Loglik	Df	AIC
Accelerated Failure Time	Exp	299	-672.54	13	1346.08
	Weibull	299	-502.444	13	1034.888
	Log-Logistic	299	-496.11	15	1026
	Lognormal	299	-47.8197	15	1022.23

This section evaluates the performance of Accelerated Failure Time (AFT) models, specifically the Weibull, Exponential, Log-logistic, and Log-normal distributions. The model selection process is based on criteria such as the Likelihood Ratio Test (LRT) and the Akaike Information Criterion (AIC). The best-fitting model, determined through these metrics, was chosen for interpretation.

The Akaike Information Criterion (AIC) and Log-Likelihood were employed to identify the most suitable Accelerated Failure-Time (AFT) model. Among the models evaluated, the Log-Normal AFT model was selected for its superior performance, evidenced by the lowest AIC and highest Log-Likelihood values, as shown in Table 1. This model was subsequently used to determine the predictors of cardiovascular disease (CVD) patient outcomes.

Table 2: Present Estimate from LogNormalAFT

	Coef	exp(coef)	se(coef)	coef lower 95%	coef upper 95%	exp(coef) lower 95%	exp(coef) upper 95%	cmp to	Z	P	-log2(p)
Alcohol usage_yes	-1.63	0.20	0.20	-2.02	-1.23	0.13	0.29	0.00	-8.12	<0.005	50.88
Body mass index_over-weight	-1.27	0.28	0.17	-1.61	-0.93	0.20	0.39	0.00	-7.32	<0.005	41.92
Diabetes militus_yes	0.98	2.66	0.17	0.64	1.32	1.89	3.74	0.00	5.61	<0.005	25.53
Pulse rate_irregular	-1.55	0.21	0.20	-1.95	-1.16	0.14	0.32	0.00	-7.69	<0.005	45.91
Region_kaduna central	-0.49	0.61	0.72	-1.90	0.92	0.15	2.51	0.00	-0.68	0.04	1.01
Region_kaduna north	0.41	1.51	0.17	0.07	0.75	1.07	2.13	0.00	2.37	0.02	5.81
Region_kaduna south	0.30	1.35	0.19	-0.06	0.67	0.94	1.94	0.00	1.62	0.10	3.26
Age	-0.01	0.99	0.01	-0.02	0.00	0.98	1.00	0.00	-1.48	0.04	2.86
Anaemia	0.04	1.04	0.15	-0.26	0.34	0.77	1.40	0.00	0.24	0.81	0.30
ejection_fraction	-0.00	1.00	0.01	-0.01	0.01	0.99	1.01	0.00	-0.17	0.87	0.20
high_blood_pressure	-0.39	0.68	0.15	-0.69	-0.10	0.50	0.91	0.00	-2.61	0.01	6.78
serum_creatinine	-0.04	0.96	0.05	-0.15	0.06	0.86	1.06	0.00	-0.81	0.42	1.26
serum_sodium	0.02	1.02	0.02	-0.01	0.05	0.99	1.05	0.00	1.28	0.20	2.32
Sex	0.02	1.02	0.17	-0.32	0.36	0.73	1.43	0.00	0.11	0.91	0.13
Smoking	0.53	1.69	0.18	0.17	0.88	1.19	2.40	0.00	2.94	<0.005	8.24
Intercept	4.82	124.23	2.04	0.82	8.83	2.27	6813.86	0.00	2.36	0.02	5.77

The table 2 is result fitted from LogNormal accelerated failure time model as some of the predictors variables such as alcoholic usage, high blood pressure, over body mass index, irregular pulse rate and many other are significant at 0.05 level while others are not which will be discuss in details in comparison.

The results of LogNormal AFT model presented in Table 2 showed that explanatory variables, region, smoking, high blood pressure, body mass index, irregular pulse rate, alcohol use , Diabetes mellitus and age have significant effect on survival of CVDs patients at 5% levels of significance. Result from LogNormal AFT modeling shown as follows: A unit increase in covariate indicates that the ean/median survival time will change by a factor of exp(coefficient).

Coefficients (coef): These values represent the log-linear change in the log-transformed survival time for a one-unit change in the predictor variable, holding all other variables constant. In other words, it tells you the direction (positive or negative) and the magnitude of the effect of each predictor on the survival time.

Exponentiated Coefficients (exp(coef)): These values are the exponential of the coefficients and can be interpreted as the multiplicative effect on the survival time. They represent how much the survival time is expected to change when the predictor variable increases by one unit.

From Log-Normal AFT regression model; Alcohol usage yes has coefficient of about -1.63. Note that the higher hazard means more at risk of the event occurring. Here, the value of exp(-1.63) is called the hazard ratio. It shows that, the coefficient of -1.63 suggests that for the corresponding predictor variable, as it increases by one unit, the log-transformed survival time decreases by approximately 1.63 units. **0.20**: The exponentiated coefficient (exp(coef)) is 0.20, which means that a one-unit increase in the predictor variable is associated with a 0.20 (or 20%) decrease in the survival time. This implies that as this variable increases, the survival time is expected to decrease by 80%. Body Mass Index_Over-Weight has Coefficient -1.27 and Exponentiated Coefficient be 0.28, this interpret that, Patients classified as overweight (Body Mass Index_Over-Weight) have a lower expected survival time (exp(coef) = 0.28) compared to those with normal BMI. Diabetes Militus_Yes has Coefficient of 0.98 and Exponentiated Coefficient to be 2.66, This indicate that, the risk (rate) of dying is 2.66 times for Patients with diabetes (Diabetes Militus_Yes) as (exp(coef) = 2.66) compared to those without diabetes.

Pulse Rate_Irregular has Coefficient of -1.55 and Exponentiated Coefficient: 0.21 which indicated that Patients with an irregular pulse rate (Pulse Rate_Irregular) have a lower expected survival time (exp(coef) = 0.21) compared to those with a regular pulse rate in the case of region patients that come from Kaduna Central has Coefficient of -0.49 and the Exponentiated Coefficient: 0.61 which indicated that Patients from Kaduna Central (Region_Kaduna Central) have a lower expected survival time (exp(coef) = 0.61) compared to the other two regions. Region_Kaduna North has Coefficient of 0.41 and Exponentiated Coefficient: 1.51 which indicated that Patients from Kaduna North (Region_Kaduna North) have a higher expected survival time (exp(coef) = 1.51) compare to Kaduna central. Region_Kaduna South has Coefficient of 0.30 and Exponentiated Coefficient to be 1.35 this also indicated that Patients from Kaduna South (Region_Kaduna South) have a higher expected survival time (exp(coef) = 1.35) compared to both Kaduna central and Kaduna north. AGE has Coefficient of -0.01 and Exponentiated Coefficient to be 0.99, this indicated that, one-year increase in age is associated with a very slight decrease in expected survival time (exp(coef) = 0.99). Anaemia has Coefficient of 0.04 and Exponentiated Coefficient be1.04, this indicated that,Patients with anaemia have a slightly higher expected survival time (exp(coef) = 1.04) compared to those without anaemia.Ejection_Fraction has Coefficient of -0.00 and Exponentiated Coefficient to be 1.00 this indicted that, the ejection fraction does not significantly affect the expected survival time (exp(coef) is approximately 1.00).

High_Blood_Pressure has Coefficient is -0.39 and Exponentiated Coefficient to be 0.68 this indicated that, Patients with high blood pressure have a lower expected survival time (exp(coef) = 0.68) compared to those without high blood pressure. Serum_Creatinine has Coefficient of 0.02 and Exponentiated Coefficient to be 1.02, this indicated that one-unit increase in serum sodium is associated with a very slight increase in expected survival time (exp(coef) = 1.02). Serum_Sodium has Coefficient of 0.02 and Exponentiated Coefficient to be 1.02, this indicated that one-unit increase in serum sodium is associated with a very slight increase in expected survival time (exp(coef) = 1.02). Sex has Coefficient of 0.02 and Exponentiated Coefficient to be 1.02 this interpret that, there is a very slight difference in expected survival time between the two sexes (exp(coef) = 1.02). Smoking has Coefficient is 0.53 and the Exponentiated Coefficient to be 1.69, this indicates that, the risk (rate) of dying is 1.69 times for patients who smoke (exp(coef) = 1.69) compared to non-smokers.

4.1 Tests for Heteroscedasticity

We analyzed the plot of estimated residuals against predicted values from the weighted least squares (WLS) method, as presented in Figure 2. The results indicate heteroscedasticity, violating the assumption of constant variance required by classical AFT models.

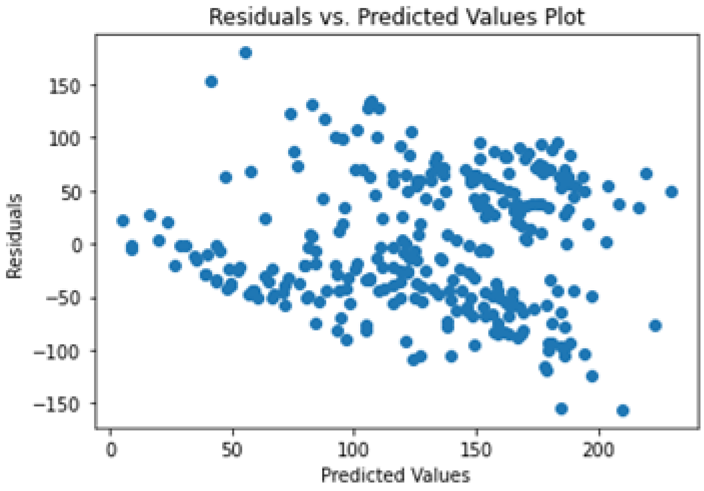

Figure 2: Visual Inspection

The scatterplot reveals a systematic widening of the spread of points as we move along the predicted or independent variables, a clear indication of heteroscedasticity. Additionally, the residuals are clustered at both ends of the predicted value range. This pattern highlights a non-constant variance of errors across the predicted values, violating the assumption of homoscedasticity. Such a violation can potentially result in biased parameter estimates and unreliable model performance.

We conducted the Breusch-Pagan test to confirm the violation of homoscedasticity observed through visual inspection (as shown in Figure 2). The hypotheses for the test are stated as follows:

H_0: The data exhibits homoscedasticity.
H_1: The data does not exhibit homoscedasticity.

The p-value obtained from the Breusch-Pagan test is 8.840728359360559e-23, which is significantly less than the 0.05 threshold. Therefore, we reject the null hypothesis (**H_0**) and conclude that the data does not exhibit homoscedasticity. This indicates the presence of heteroscedasticity, consistent with the observations in Figure 2.

Table 3: Weighted least square estimation

	coef	std err	T	P>|t|	[0.025	0.975]
Const	1.4554	0.958	1.519	0.130	- 0.431	3.342
Age	-0.0087	0.001	-10.167	0.000	-0.010	-0.007
Anaemia	0.4365	0.019	22.636	0.000	0.399	0.474
ejection_fraction	-0.0022	0.001	-3.361	0.001	-0.003	-0.001
high_blood_pressure	-0.1280	0.017	-7.447	0.000	-0.162	-0.094
serum_creatinine	0.0112	0.004	2.500	0.013	0.002	0.020
serum_sodium	0.0253	0.002	12.646	0.000	0.021	0.029
Smoking	0.1200	0.027	4.389	0.000	0.066	0.174
Sex	-0.3769	0.033	-11.410	0.000	-0.442	-0.312
REGION_KADUNA NORTH	0.1597	0.027	5.951	0.000	0.107	0.213
REGION_KADUNA CENTRAL	-0.3990	0.030	-13.164	0.000	-0.459	-0.339
REGION_KADUNA SOUTH	0.0310	0.032	0.965	0.335	-0.032	0.094
ALCOHOL USAGE_YES	0.1726	0.056	3.106	0.002	0.063	0.282
BODY MASS INDEX_OVER-WEIGHT	-0.4649	0.121	-3.828	0.000	-0.704	-0.226
PULSE RATE_IRREGULAR	-0.4649	0.926	-2.038	0.042	-3.708	-0.064
DIABETES MILITUS_YES	0.2772	0.026	10.660	0.000	0.226	0.328

The analysis revealed the following findings for the coefficients of the independent variables:

- **Age**: The coefficient is −0.0087, indicating that for every additional year of age, the dependent variable decreases by 0.0087, holding other variables constant. This effect is statistically significant (p=0.000).
- **Anemia**: The coefficient is 0.4365, suggesting that having anemia increases the dependent variable by 0.4365 compared to not having anemia, holding other variables constant. This effect is statistically significant (p=0.000).
- **Ejection Fraction**: The coefficient is −0.0022. For each unit increase in ejection fraction, the dependent variable decreases by 0.0022. This effect is statistically significant (p=0.001).
- **High Blood Pressure**: The coefficient is −0.1280. Having high blood pressure reduces the dependent variable by 0.1280 compared to not having high blood pressure. This is statistically significant (p=0.000).
- **Serum Creatinine**: The coefficient is 0.0112. For each unit increase in serum creatinine, the dependent variable increases by 0.0112. This effect is statistically significant (p=0.013).

- **Serum Sodium**: The coefficient is 0.0253. For each unit increase in serum sodium, the dependent variable increases by 0.0253. This is statistically significant (p=0.000).
- **Smoking**: The coefficient is 0.1200. Smokers have a 0.1200 higher dependent variable compared to non-smokers. This effect is statistically significant (p=0.000).
- **Sex**: The coefficient is −0.3769. Being male reduces the dependent variable by 0.3769 compared to females, holding other variables constant. This effect is statistically significant (p=0.000).
- **Region**:
  - Kaduna North: The coefficient is 0.1597, indicating that patients in Kaduna North have higher dependent variable values compared to the reference region (p=0.000).
  - Kaduna Central: The coefficient is −0.3990, showing lower values for patients in Kaduna Central compared to the reference region (p=0.000).
  - Kaduna South: The coefficient is 0.0310, which is not statistically significant (p=0.335).
- **Alcohol Usage**: The coefficient is 0.1726. Alcohol users have a 0.1726 higher dependent variable compared to non-users. This effect is statistically significant (p=0.002).
- **Body Mass Index (Overweight)**: The coefficient is −0.4649. Being overweight reduces the dependent variable by 0.4649 compared to having a normal weight. This is statistically significant (p=0.000).
- **Pulse Rate (Irregular)**: The coefficient is −0.4649. Having an irregular pulse rate reduces the dependent variable by 0.4649. This effect is statistically significant (p=0.042).
- **Diabetes Mellitus**: The coefficient is 0.2772. Patients with diabetes mellitus have a 0.2772 higher dependent variable compared to those without diabetes. This effect is statistically significant (p=0.000).

Nevertheless, we compared the performance of LNAFT with the WLSE since the LNAFT was the best among the classical methods. The results in Table 3 revealed that the WLSE is more accurate than the LNAFT. It resulted in efficient estimates of covariates effect on cardiovascular patients' survival since the narrow CI indicates a relatively small standard error and AIC 712.08. It also identified more significant covariates since serum-creatinine, ejection-fraction,anamia and seum-sodium were additional significant covariates that were not identified by LNAFT as we considered the same set of covariates in all models at the beginning, we also considered all the covariates in the final models as per their ability to identify significant effects.

5. CONCLUSION

We utilized AFT models based on the classical and WLSE methods on a cardiovascular dataset we compiled among possible parametric AFT models, the lognormal AFT model fitted the data well. We compared the results from this model with the result from WLSE. The width of confidence intervals from the WLSE was found to be shorter than that of the classical methods. The WLSE also detected more significant covariates. As a result, the WLSE performed best in providing more significant effects and precise estimates. However, the data was heteroscedastic. Thus, we recommend future researchers extend the application of WLSE to a homoscedastic real dataset with more covariates to ascertain its validity. They should utilize WLSE rather than the standard AFT methods when the homoscedasticity assumption is violated to obtain efficient estimates. Moreover, health workers should be more cautious when a patient is in advanced clinical stages, old in age, over weight body mass index, have diabetes mellitus, have irregular pulse rate, take alcohol and as well smoked.

REFERENCES

Abtsega, S., Ayalew, M., Abiso, E. & Kabtamu, T. G. (2019). Survival Analysis of Factor Affects Survival Time of Hypertension Patients. *Open Journal of Modelling and Simulation*, 7, 177-189.

Alvaro, F. G. & René, B. (2021) Survival analysis of patients with heart failure in the Ecuadorian Andean population. *Medwave*;21(07);84-94.

Austin, P. C., Steyerberg, E. W., & Putter, H. (2021). Fine- Gray subdistribution hazard models to simultaneously estimate the absolute risk of different event types: Cumulative total failure probability may exceed 1. *Statistics in Medicine*, 40(19), 4200–4212.

Barnwal, A., Cho, H., & Hocking, T. (2022). Survival regression with accelerated failure time model in XGBoost. *Journal of Computational and Graphical Statistics*, 31(4), 1292–1302.

Belaynesh, Y.E. & Zeytu, G.A. (2021). Comparison of survival models and assessment of risk factors for survival of cardiovascular patients at Addis Ababa Cardiac Center, Ethiopia: a retrospective study. *National Library of Medicine. National Centre of Biotechnology Information*.

Edward, H., Achmad, E., Geert, M. & Alain, G. B. (2015). Comparison of risks of cardiovascular events in the elderly using standard survival analysis and multiple-events and recurrent-events methods.*BMC Medical Research Methodology*. 15(15); 156-167.

Fox, G. A. (2020). Failure-time analysis: Emergence, flowering, survivorship, and other waiting times. In Design and analysis of ecological experiments (pp. 253–289). *Chapman and Hall/CRC.*

Hui, P. Z., Xin, X., Chuan, H. Y., Ahmed, A., Shun, F. L., Yu, K. D. (2011). Application of Weibull model for survival of patients with gastric cancer. *BMC Gastroenterology 2011, 11:1 http://www.biomedcentral. com/ 1471-230X/11/1*.

Javeria K.,Muhammad A., & Zahra A.(2021). Influence Diagnostics in Log-Normal Regression Model with censored data. *Mathematical problem in engineering*,2021,02-15.

Orbe, J., & Núñez-Antón, V. (2006). Alternative approaches to study lifetime data under different scenarios: From the PH to the modified semiparametric AFT model. Computational Statistics & Data Analysis, 50(6), 1565–158.

Piotr, M. S. & Aaron, J. M. (2022). Scalable algorithms for semiparametric accelerated failure time models in high dimensions. *Department of Statistics and Genetics Institute University of Florida, Gainesville, FL*.

Qingkai D., Binxia & L., Hui Z.(2023) Weighted least squares model averaging for accelerated failure time models.*scienceDirect*,184,02-20.

Rana A. W. & Shuaib, M.K. (2006). Comparative Distributions of Hazard Modeling Analysis. *Pak. j. stat. oper. res.* L(2); 127-134.

Richard L. (2012).Survival Analysis models and Applications. *First Edition, John Wiley & Sons Ltd, The Atrium, Southern Gate, Chichester, United Kingdom*

Spoto, B., D’Arrigo, G., Tripepi, G., Bolignano, D. & Zoccali, C. (2021). “Serum gamma-glutamyltransferase, *oxidized LDL and mortality in the elderly,” Aging Clinical and Experimental Research*, 33(5); 1393–1397.

Swindle, R. and Mukhopadhyay, K. (2011). “Survival analysis in clinical trials: basics and must know areas,” *Perspectives in Clinical Research*, 2(4); 145–148.

Tanvir, A., Assia ,M., Sajjad ,H. B., Muhammad ,A., Muhammad ,A. R. (2017) Survival analysis of heart failure patients: *A case study. PLoS ONE* 12(7); 567-603.

Victor, M. O., Ezekiel, U. N., Ifeoma, I. U., Adeseye, A. A., Ekene, E. C., Phillip, T. B., Ross, S. R. & Timothy, C. S. (2017). Cardiovascular disease risk factors in a Nigerian population with impaired fasting blood glucose level and diabetes mellitus. *BMC Public Health*.

Yesuf, A. M. & Ding-Geng, C. (2021). Accelerated failure-time model with weighted least-squares estimation: application on survival of HIV positives. *Archives of Public Health*, 9(3); 79-88.

Yu, L., Liu, L. (2013). Weighted least-squares method for right-censored data in accelerated failure time model. *Int Biom Soc*. 69(2); 58–65.

Yu, L., Liu, L. (2018). A homoscedasticity test for the accelerated failure time model. *Comput Stat*

Zhongheng, Z. (2016). Parametric regression model for survival data: Weibull regression model as an example. *Annals of Translational Medicine*;4(24);48-56.